Phonology Modelling for Expressive Speech Synthesis: a Review

Raheel Qader, Gwénolé Lecorvé, Damien Lolive, Pascale Sébillot

Publications Internes de l'IRISA
ISSN : 2102-6327
PI 2020 –

Phonology Modelling For Expressive Speech Synthesis: a Review

Raheel Qader[*] , Gwénolé Lecorvé[*] , Damien Lolive[*] , Pascale Sébillot[**]
raheel.qader@irisa.fr, gwenole.lecorve@irisa.fr, damien.lolive@irisa.fr, pascale.sebillot@irisa.fr

Abstract: Expressive speech processing is an important scientific problem as expressivity introduces a lot of variability into speech. This variability leads to a degradation of speech application performances. Variations are reflected in the linguistic, phonological and acoustic sides of speech. However our main interest is on phonology, more precisely the study of pronunciation and of disfluencies. Both of these fields have huge impacts on speech.

This report is a bibliographical review of the state of the art in expressivity and phonology modelling. Although the main focus will be on speech synthesis, we will discuss works about automatic speech recognition as well because expressivity modelling in phonology is a cross-domain problem.

Key-words: Phonology, pronunciation, disfluencies, expressivity, speech synthesis, automatic speech recognition.

Modélisation de la phonologie pour la synthèse de parole expressive

Résumé : *L'expressivité introduit beaucoup de variabilité dans la parole. Cette variabilité touche des aspects aussi linguistiques, phonologiques qu'acoustique et conduit généralement à des dégradation des applications de traitement de la parole. Ainsi, le traitement de la parole expressive est un problème important. Précisément, notre intérêt principal se porte sur l'étude la phonologie, plus précisément celle de la prononciation et des disfluences, ces deux champs ayant chacun un rôle considérable dans la parole.*

Ce rapport est une étude bibliographique des travaux liées à l'expressivité et à la modélisation de la phonologie. Le cadre de cette étude est principalement celui de la synthèse de la parole. Néanmois, comme la modélisation phonologique de l'expressivité est une problématique multi-domaine, nous aborderons également des travaux issus du monde de la reconnaissance automatique de la parole.

Mots clés : *Phonologie, prononciation, disfluences, expressivité, synthèse de la parole, reconnaissance automatique de la parole.*

---

[*] Équipe Expression.
[**] Équipe LinkMedia commune avec INRIA.

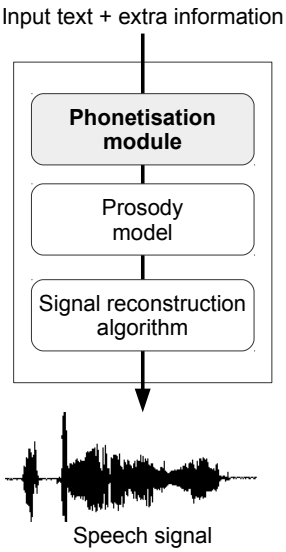
Contents

1 Introduction

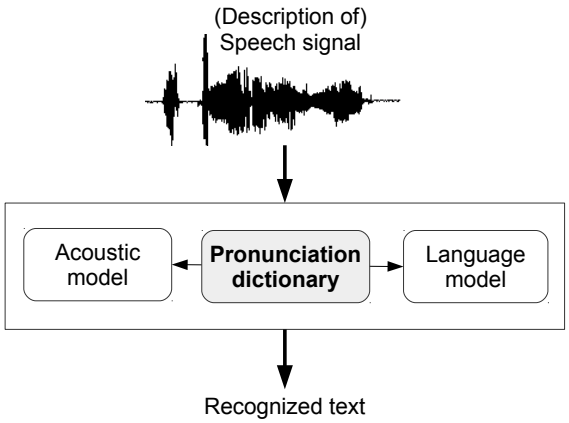
Human speech has two different distinctive aspects, the linguistic content which is conveyed from a speaker to another and the style in which the speech is uttered [Eide et al., 2004]. Whereas the linguistic content is rather to bring direct information about the meaning, the impact of style is more implicit. In the recent years, in speech processing, style has been particularly studied through the spectrum of *expressivity*, that is the expression of different states of mind like emotions, so-called speaking styles (casual, formal, etc.) or intentions. Speech data exhibiting such characteristics is referred to as expressive speech, as opposed to neutral speech where style is absent.

Expressive speech processing is an important scientific problem as expressivity introduces a lot of variability into speech. In linguistics, expressivity may change the choice of words or of syntactic structures. In acoustics, it impacts various characteristics like the energy, pitch, duration, etc. In this PhD, we rather focus on the changes in *phonology*. Generally speaking, phonology is the study of sounds. In the context of this work, it will more precisely gather the study of pronunciation and of disfluencies. Variabilities in pronunciation can appear through deletion, insertion or substitution of phones compared to standard pronunciation. For instance, the standard pronunciation of the word *and* is /ænd/ in English, but in spontaneous speech it is mostly pronounced as /æn/, the final phone /d/ being dropped. Regarding disfluencies, expressive speech introduces phenomena like filled pauses, repetitions, etc. For instance, in unprepared speech, speakers tend to overuse filled pauses [Corley and Stewart, 2008], mainly because they are still planning the flow of their discourse while speaking.

These variabilities introduce many problems in speech applications and lead to a degradation in their performance. In this work, we are specifically interested in Text-To-Speech (TTS) with some extensions to Automatic Speech Recognition (ASR). An overview of how TTS and ASR work



(a) TTS



(b) ASR

Figure 1: Architecture of (a) TTS and (b) ASR systems

in given in Figure 1. TTS systems convert a written text into a corresponding intelligible and natural speech signal. First, the raw input text is processed by many modules and enriched with various features like part of speech, syntactic analysis, etc. Based on this information a phonetic representation of the text is computed. This representation is completed with prosodic information (mainly intonation). Finally, a speech reconstruction algorithm generates the output speech signal. Whereas current systems work well in generating neutral and intelligible speech, they still lack a modelling of expressivity. More precisely, the phonetisation models are usually static and unable take into account variabilities due to expressivity. Worse, current TTS systems do not integrate any disfluency model. On the other hand, ASR systems work in an opposite way. Their goal is to accept a speech signal as input and output the text uttered in the signal. While an ASR system does not consider the actual signal but rather a parametrisation of it, the main work is achieved by three components. The acoustic model aims at modelling each elementary sound. The language model provides probabilities of word sequences in the considered language. As a bridge between those two models, the pronunciation dictionary links each word to its phonetic representation. Among the various problems faced by these components, phonology-related problems mainly concern the pronunciation dictionary, and in a lesser extent the language model. Mainly the problem arises when the speaker pronounces words in an unusual way or words which are absent from the dictionary, the latter being called Out-Of-Vocabulary words (OOVs). Regarding disfluencies, the problem is to consider their probability to appear in the middle of word sequences.

The purpose of this PhD is to study phonology modelling in the frame of expressive speech synthesis. The following is a list of the most important objectives that we will try to achieve during this PhD:

- Extracting and analysing characteristics of expressivity in speech corpora.
- Modelling phonology (pronunciation and disfluencies) of expressive speech, with an emphasis on statistical/probabilistic approaches.
- Integrating expressive phonological models in a TTS system to generate expressive speech.

In addition, this PhD focuses on the symbolic representations of speech. Hence signal processing aspects will mainly be left aside. Instead, extensive usage of Natural Language Processing (NLP) techniques will be favoured. Finally, within the course of this PhD, major efforts will be spent on studying pronunciation modelling, while the problem of disfluencies is considered as more exploratory.

This document seeks to review the state of the art in expressivity and phonology modelling. Although the main focus is on speech synthesis, expressivity modelling in phonology is a cross-domain problem and this bibliographical study will thus discuss works spanning over both TTS and ASR. The remainder of this report is as follows. In Section 2, expressivity is defined and its types are explained. Sections 3 and 4 then review the literature about pronunciation and disfluencies respectively. Finally in Section 5 a discussion about the work that we have done and the work that we are planning to do in the rest of the PhD is presented.

2 Expressivity

Expressivity is a complex concept. [Govind and Prasanna, 2013] defines it as the vocal indicator of various emotional states that is reflected in the speech waveforms. More generally, emotional states can be extended to psychological characteristics of a speaker at a giving time, e.g., emotions, speaking style, personality, intention, etc. Expressivity makes human speech richer and more natural by adding an extra layer of non-linguistic information to the speech. A linguistic message can convey different meanings based on the expresivity factors that are added to it such as emotion, and speaking style. For instance, a linguistic message that is uttered with a high pitch and intonation at the end might be understood as a question, while the same utterance with a neutral pitch will be considered as normal phrase. According to [Govind and Prasanna, 2009], expressivity is considered to be the basic block of effective communication as it makes a conversation more effective and increase the involvement of listeners.

Emotion, speaking style and intention are all known to affect phonology, i.e. pronunciation and disfluencies. In this PhD, we will be mostly concentrating on different speaking styles and accents as a specific case of speaking styles, without considering emotions. Nevertheless, emotions will be still discussed in the following section as it is an important part of expressivity. These aspects are defined and reviewed in this section.

2.1 Emotion

Emotions can be expressed in different forms. They can generally be labelled and categorized into positive and negative emotions. Positive emotions include joy, pride, love, relief, hope, compassion while negative ones include anger, anxiety, guilt, shame, sadness, envy, jealousy, and disgust [Adda-Decker et al., 2005]. In a finer way, emotions can be represented as points of a continuous space. Especially, [Russell, 1980] suggests a 2-dimensional representation of emotions where the first dimension stands for pleasure-displeasure and the second for the arousal degree. Figure 2 places most known emotions in this space based on their characteristics. The horizontal axis is for pleasure while the vertical axis is for the arousal degree. For example, anger can be defined as a moderate displeasure (left side) and a neutral arousal (center of the axis). Using this representation, the closeness between emotions can be computed.

Studies have shown that speech and emotion have a very strong correlation and that emotion plays a critical role in communication [Iida et al., 2003]. Emotion might affect a speaker's choice of words i.e. the speaker mostly utters the type of words that reflect his/her emotional state. However, emotions are more tightly bound to acoustic characteristics rather than to linguistic ones. The acoustic parameters that are linked to emotions consist of fundamental frequency, formant frequencies, intensity and duration [Schröder, 2009]. In TTS, availability of these parameters is necessary for the production and identification of emotion types. In ASR, the variability of these parameters based on a speaker's emotional state brings recognition rate drops compared to neutral speech [Polzin and Waibel, 1998]. As the aim of this work is to study phonological variations of expressive speech using NLP techniques, emotions will be less studied in this PhD.

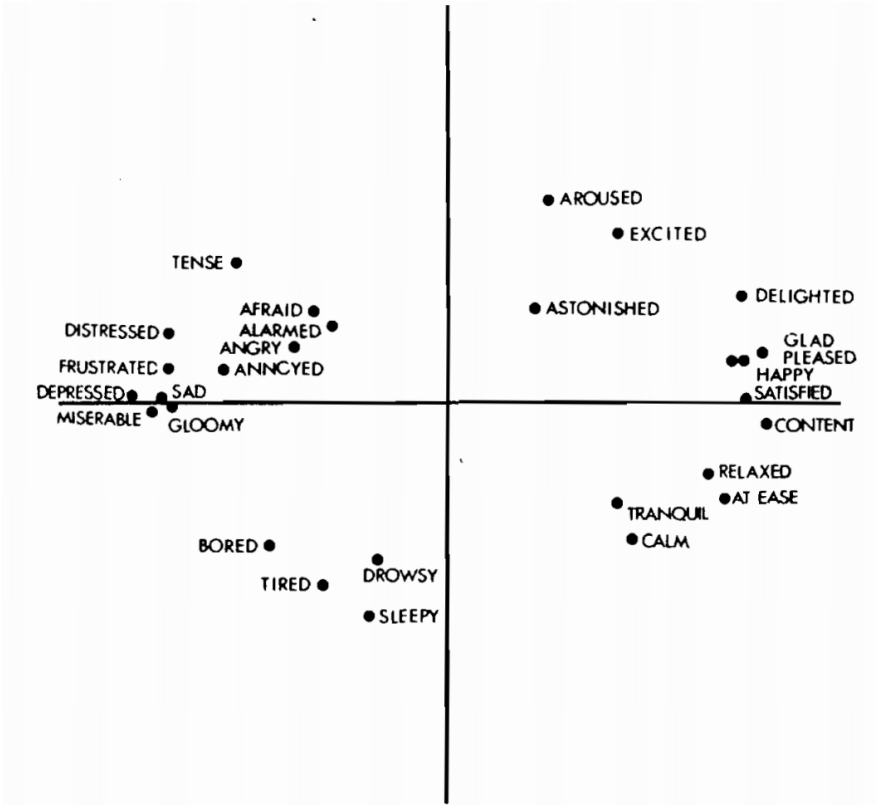


Figure 2: 2-dimensional representation of emotions (source: [Russell, 1980])

2.2 Speaking style

The notion of "speaking style" is still ambiguous to most researchers, as speaking style not only varies from region to region but also from one person to another. According to [Parlikar, 2013], the same speaker can adopt many different speaking styles based on the performed task. Speaking style is most likely to change based on the environment that the speaker is found in. For instance, a speaker usually has a relaxed speaking style when talking to a friend in an informal conversation, while the same speaker is likely to change his style of speech into a formal one during a corporate meeting. Different speaking styles can mostly be observed in spontaneous/casual speech, in which speakers prefer to communicate in their own style.

Several factors lead to the generation of different speaking styles among speakers. These factors include acoustic and phonological changes. Acoustic variations include intonation, duration, fundamental frequency, intensity and loudness [Laan, 1997] while phonological changes, as suggested by [Adda-Decker and Lamel, 1999], can be related to a variety of factors such as rate of speech, syllabic structure of words, individual speaker habits and regional dialect. These phonological changes occur at the phonemic level, for instance, insertion, deletion and substitution of phones. According to [Laver, 1994] (cited by [Strik and Cucchiarini, 1999]), changes occur as the degree of formality becomes less. In this study, we will look into to the phonological changes in depth.

2.3 Accents

Accent can be thought as a particular case of speaking style. As such, it takes part in expressivity in general. Both native and foreign speakers of a language seem to have a specific accent. Native speakers are affected by regional accents while foreign speakers are affected by the patterns which they carry from their own language. As reported by [Arslan and Hansen, 1996], foreign speakers can be identified based on the appearance in their speech of certain patterns which cannot be found in the speech of native speakers. Foreign speakers who have acquired

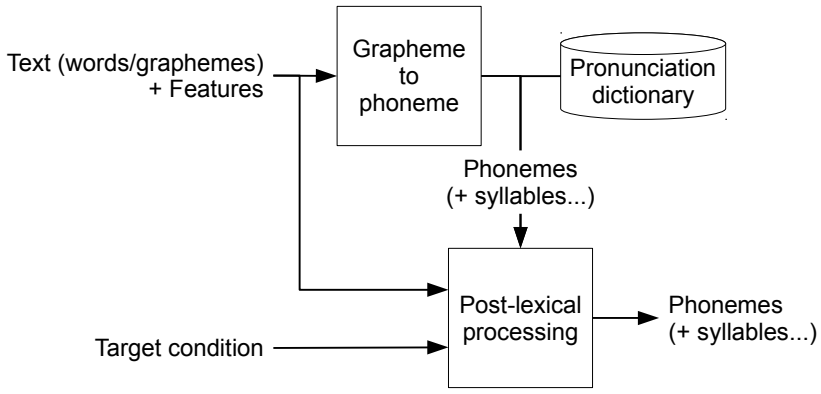


Figure 3: Overview of the main modules and data involved in pronunciation modelling

the language in an early age are also reported to be able to minimize their accent. Moreover, [Arslan and Hansen, 1996] defines a foreign accent as the patterns of pronunciation features which characterize an individual's speech as belonging to a particular language group. In this research, we are rather interested in accents of foreign speakers in the first place, as it is believed that patterns of foreign speakers are more obvious and easier to detect than those of native speakers.

In short, expressivity is an important characteristic of human speech that differentiates speech of individuals from each other and makes speech more natural. However this usually leads to poor performances of speech applications. In the next section, some of the major concepts and studies within the area of expressivity and pronunciation modelling will be reviewed.

3 Pronunciation modelling and generation

In general, pronunciation modelling is a way to link the orthographic representation of written words with their corresponding sequence of elementary sounds to be spoken. These elementary sounds are referred to as *phonemes*. In the frame of speech synthesis, the main usage of a pronunciation model is to generate these phonemes based on an input text. Seeking to integrate expressivity in TTS, we are especially interested in building pronunciation models that can learn or adapt to specific pronunciation patterns and variations. Nonetheless, as already mentioned in introduction, pronunciation modelling is also a problem in ASR. This review will thus not be limited to speech synthesis.

In this section, we first introduce the general concepts and tasks in pronunciation modelling before explaining the evaluation methodology, then we present the important features to address these tasks. Finally, a review of the major techniques and results from the literature are given.

3.1 Overview

Pronunciation modelling can be decomposed into two tasks with their own models. This decomposition is illustrated in Figure 3. The input of pronunciation models are usually words or graphemes and the output are phonemic representations. In sophisticated approaches, extra features can be considered such as part of speech, etymology, etc. Likewise output phonemes can come along with syllables, information about stress, etc. In more details, the first task consists in converting the input text into its corresponding phonemic representation as it should be considered in the standard language. This phonemic representation will be referred to as *canonical* pronunciation in the rest of the document. Alternatively the output can be the list or the graph of all possible pronunciations without explicitly determining the one to be chosen. Producing canonical pronunciations is achieved by either searching in a hand-crafted dictionary, or using a Grapheme-To-Phoneme (G2P) converter. The second task in pronunciation modelling is to modify or rerank the canonical pronunciations in

order to reflect a a specific target condition e.g. accented speech, emotion, etc. This task is known as post-lexical processing.

3.2 Evaluation methodology

Pronunciation models can be evaluated based on both the generated signal and the generated phonemic representations. As we are more interested in the phonemic level changes in this PhD, we will only consider the evaluation methods and measures for the linguistic side. On the phonemic level, two metrics are mostly used: Phone Error Rate (PER) and Word Error Rate (WER). These metrics are useful to compare the output of a given pronunciation model to a baseline or competing methods [Strik and Cucchiarini, 1999]. In order to compute these two measures, the generated sequence of phonemes should first be aligned with the phonemes of a reference pronunciation. Each phoneme is aligned with one or zero phoneme of the other sequence. One of the most widely used string alignment algorithms is the Levenshtein algorithm. This algorithm calculates the number of insertion, deletion and substitution operations needed to transform the reference sequence into the generated one. This determines the so-called edition distance between them [Wieling et al., 2009]. The result of the alignment process then reveals the phonemes and words which are predicted correctly. Using this information, we can calculate the PER by dividing the number of mispredicted phonemes over the total number of phonemes N:

$$PER = \frac{S + D + I}{N} \ , \tag{1}$$

where S, D, and I are the numbers of substitutions, deletions and insertions respectively. When different references exist for a same word, only the one with the lowest edition distance is used for PER computation. Calculating WER is almost the same, however words are be taken into account instead of phonemes. WER is frequently higher than PER on a given test set. This is because a word is considered as mispredicted as soon as there is at least one phoneme error.

Besides these two measures, a very similar measure called *Accuracy* can be used to measure performance of a system. Accuracy measures the correctly predicted units as opposed to error rates [Lopes and Perdigão, 2011]. For instance, the equation for phoneme accuracy is as follows:

$$\textit{Phoneme accuracy} = \frac{N + S + D + I}{N} = 1 - PER \ . \tag{2}$$

As for PER and WER, accuracies are expressed as percentages.

3.3 Useful features

Pronunciation models, as described previously, handle various types of information as input and output. It is very important to understand these features and to study how they are used within the context of pronunciation modelling. This section describes some of the main features which can be derived from phonemes, syllables, speech signals and linguistics.

3.3.1 Phonemes and phonetic features

Sound units can be divided into two basic types: phones and phonemes. *Phones* are the basic units which describe any single speech sound with a single articulatory configuration [Taylor, 2009]. Phones are mostly related to the acoustic features found in a particular spoken language, disregarding the phonological role in the language. *Phonemes* are considered as a more abstract layer of sounds that can be used to represent the linguistic constituents of a language. A single phoneme can represent several phones. For example, in English, the phoneme /p/ has more than one realization such as in the word *pit* and *spit*. The /p/ in *pit* is more aspirated meaning that more air is expelled at the end of the sound [Fosler-Lussier, 1999c]. These two different realizations are called *allophones* of /p/. Allophones can be thought as a link between phones and phonemes. In this work, we will

mostly use phonemes because modelling pronunciation is usually conducted by exploiting linguistic features rather than acoustic features.

Researcher and linguists working in the area of phonology use different system for codifying sounds. One of the most widely used one is called International Phonetic Alphabet (IPA) [International Phonetic Association, 1999]. In IPA, each sound is represented with a symbol and all the sound in any language can be represented using these IPA symbols. Moreover, symbols can be specialized by adding them characteristics (for instance, to define a phone as stressed or long). This is done by adding special marks called *diacritics*. A diacritic is a mark added to letters which slightly changes the sound of the letter. For instance, in the French word *bonjour*, the graphemes *on* are pronounced as /õ/, where /ɔ/ is the phoneme and ∼ is the diacritic for nasalization. Having such a standard system of sound representation makes easier working and sharing ideas in this area.

Several types of information can be extracted from phonemes to be used as features in pronunciation models. For instance phonemes can be categorized into vowels and consonants, voiced and unvoiced[1], nasalized or not, etc.

3.3.2 Syllables and syllabic features

Contrary to what most people think, syllables are not just mere sequences of phones but they are completely distinguishable from phones in the sense that they have structural integrity, and are tightly connected to the higher tiers of linguistic organization [Greenberg, 1999]. The structure of syllables is divided into 3 main parts which are, from left to right, the onset, the nucleus and the coda. The conjunction of the nucleus and the coda forms the rhyme. The nucleus is mandatory and consists of a vowel, while the other two parts are optional and a made of consonants and semi-vowels. For instance, the word *kitten* is canonically pronounced /kitən/. This pronunciation is made of the 2 syllables /ki/ and /tən/, and the second syllable has the follow structure: the onset is /t/, the nucleus is /ə/ and the coda is /n/.

Syllables can also be categorized into *open* and *closed* syllables [Moats, 2004]. Open syllables end with a vowel, that is their coda is empty, like in the first syllable /ki/. At the opposite, closed syllables have phonemes for each of the three parts. Questions arise for closed syllables regarding the segmentation of a word. For example, the phoneme string /kitən/ could have also been split into the syllables /kit/ and /ən/. Deciding this problem depends on the language while it may also be disambiguated thanks to a spoken example. A syllable can also be *stressed* or *unstressed*. Stressed syllables are those with high amount of energy and longer duration. According to [Dilts, 2013], stressed syllables are less likely to be deleted during spontaneous speech. Another factor that leads to variations in speech is the syllable location within a word. As its reported in [Vazirnezhad et al., 2009], the initial and the middle syllables of words show very low ratios of deletion in spontaneous speech, whereas this ratio is higher in the final syllable. Lastly, another feature can be derived from syllable: the rate of speech (or speech rate). The rate of speech is counted as the number of linguistic units per second, most of the time syllables per second. [Fosler-Lussier and Morgan, 1999] reported that when the rate of speech is high, pronunciation of words are more likely to diverge from their canonical forms.

Several studies [Greenberg, 1999, Fosler-Lussier, 1999a, Fosler-Lussier, 1999b] have suggested that moving toward larger linguistic units such as syllable or word instead of phonemes will improve the performance of pronunciation models. The main idea here is that variations on phonemes are not systematic, leading phoneme-based models to predict too many variants for each phone.

3.3.3 Speech signal features

Using linguistic content to model pronunciation is a widely spread method. At the acoustic level, the pronunciation variants are modelled by adapting acoustic models. In TTS, these models can be used

[1] Voiced sounds are those in which the vocal folds vibrate during its articulation, while voiceless sounds are those that the vocal folds do not vibrate.

to generate different pronunciations. Acoustic models exploit different types of features to extract useful information such as F0, pitch and formants and to remove noise from the speech signal. According to [Taylor, 2009] F0 is the fundamental frequency of a speech signal which is the driving frequency of the vocal cord. Pitch can be explained as the perceived fundamental frequency after undergoing some linearities or errors in the perception. Lastly, formants as described by [Fant, 1971] (cited in [Guilleray, 2012]) are "the spectral peaks of the sound spectrum of the voice", or "an acoustic resonance of the human vocal tract". Formants are extremely important especially in identifying types of vocalic phonemes, i.e. vowels.

Acoustic models represent speech signals with feature vectors such as PLP and MFCC. As reported in [Dave, 2013], each of these feature vectors use a different representation of speech signals, for instance, MFCCs is based on the frequency domain and it uses the Mel scale which is based on the human ear scale. Whereas, in PLP, irrelevant information of the speech is discarded.

3.3.4 Linguistic features

Many linguistic features can be derived from words to be phonetized. Though, we only present the 3 most important here.

The first is the *frequency of words in language*. Indeed, it is believed that frequent words are easier for listeners to recognize and understand, while rare words are more difficult. However, in reality, speakers tend to use shorter variants for frequent words and pronounce them a bit fast. In contrast, rare words are pronounced more slowly and closer to their canonical forms. That is why frequent words are sometimes less recognizable by speech recognition applications [Vazirnezhad et al., 2009]. For instance, function words which usually comes at the top of frequent words (articles, auxiliary verbs, etc.) in most languages are mostly shortened during speech [Bell et al., 2009]. Speakers usually use shorter variants of these words which makes it more difficult to be recognized. Studies of large speech corpora have shown that the number of phonemes in frequent words is lower than the infrequent words and deletion of phonemes is higher in frequent words [Vazirnezhad et al., 2009].

Second, words can show different behaviours in speech according to their frequency inside a particular *discourse*. For example when a word is mentioned for the second time in a discourse, its duration is shorter than the first occurrence of the word [Bell et al., 2003]. According to [Bell et al., 2009], function words are also short when they are frequent and shorter when they are repeated inside a discourse.

Finally, Part-Of-Speech (POS) can be an indicative feature for pronunciation modelling. POS is the linguistic category of a word such as verb, noun, adjective, etc. [Schachter and Shopen, 1985]. POS is a critical feature of many speech and NLP applications. In some languages like French, POS is extremely important for pronunciation modelling, as it directly affects the pronunciation of some word categories. For instance, the grapheme string "ent" at the end of third person plural French verbs is silent, e.g. "*ils parl**ent***", whereas in adverbs like *vraim**ent*** it is pronounced /ɑ̃/. Moreover, POS is also used in identification of the syntactic and semantic structure of sentences.

3.4 Grapheme-to-Phoneme (G2P) conversion

According to [Van Den Bosch and Daelemans, 1993], the objective of grapheme-to-phoneme conversion is to generate a sequence of phonemes for a given word from its spelling. Thus, it generates a sequence of phonemes from a sequence of characters, these characters being called graphemes. The output of G2P can be recomputed by post-lexical processors to further improve the pronunciation or to model variations in speech. An explanation of post-lexical processing can be found later in the next section. while this section focuses on the main G2P techniques, namely knowledge-based and data-driven techniques, with a particular insight of statistical approaches.

3.4.1 Knowledge-based techniques

G2P converters can be implemented in several ways, but probably the most straightforward one is to store all the possible pronunciations in a pronunciation dictionary and then to look up in this dictionary at run time for each input word. The technique has the disadvantage of not being able to predict the pronunciations of OOVs. In a more flexible approach, rule-based techniques are based on the idea that pronunciation of a letter can be predicted from the context of that letter [Pathak and Talukdar, 2013]. Most of the old rule-based systems consisted of hand-written rules. The drawback of this technique is that it requires experts to craft the rules for each language in order to achieve a good performance. In addition, it might perform very poorly in languages such as English and French where the connection between graphemes and phonemes can be ambiguous [Deshpande, 2013].

3.4.2 Data-driven techniques

Instead of using hand-written rules, data-driven techniques seek to learn the rule automatically from examples and apply them accordingly. Neural networks and analogy learning are the most widely known machine learning techniques for data-driven G2P conversion. Neural networks is a classification approach which predicts each phoneme independently by only taking into account the context of the current grapheme [Jiampojamarn, 2011]. One of the first examples of using neural networks in bulding grapheme-to-phoneme systems is the NETtalk system [Sejnowski and Rosenberg, 1987] (cited by [Taylor, 2009]). NETtalk consisted of 3 layers. The input layer was made of 203 neurons, the hidden layer of 80 and the output 26 units coding the phoneme to be produced. It considered only seven characters at a time with a window of three characters from right and left of the target character. The input and output of the system were encoded with various features (voiced, stress, syllable boundary, etc.).

The idea behind pronunciation by analogy comes from the studies of how humans learn the pronunciation of new words. When a human is given a new word, he/she learns its pronunciation by comparing it to the nearest known words and adapting or combining their pronunciations [Taylor, 2009, Dedina and Nusbaum, 1991]. For instance, if we consider the word *fax* as our target new word, a human would automatically think of a similar word such as *tax* and adapt its pronunciation. Pronunciation by analogy algorithms work by comparing substrings of the unknown word to those extracted from the pronunciation dictionary and find the closest match [Damper and Eastmond, 1997].

3.4.3 Statistical techniques

The statistical approach is more recent compared to the other two techniques. It includes models such as Hidden Markov Models (HMM), Conditional Random Fields (CRFs) and joint n-gram models.

In HMMs, the model is allowed to use the previously predicted phonemes for the future decisions. Phonemes are represented as states of a Markov chain and linked to to their most likely corresponding graphemes [Karanasou, 2013]. The Viterbi algorithm is used to predict the optimal state sequence [Viterbi, 1967]. An example of using HMM in G2P is the work of [Taylor, 2005] in which he suggests that HMMs alone are not sufficient for G2P modelling but by adding a preprocessing step they can be improved. The preprocessing step rewrites some of the graphemes and rearrange them. For instance, a word like *hate* would be rearranged to *haet* and graphemes x were rewritten as *ks*.

On the other hand, CRFs are also probabilistic models used for labelling sequential data. CRFs are widely used in NLP to solve sequential problems such as POS tagging [Tellier et al., 2010], G2P [Illina et al., 2011, Wang and King, 2011], etc. CRFs offer several advantages over HHMs by relaxing the strong independence assumptions [Lafferty et al., 2001]. CRFs compute output probabilities $\Pr(\vec{y}|\vec{x})$ of a sequence of observations $\vec{y} = (y_1, \ldots, y_n)$ from a sequence of in-

put $\vec{x} = (x_1, \ldots, x_n)$. The most important part of CRFs are the so-called feature functions. When used in a task like G2P conversion, each feature function takes as input a grapheme that we want to convert to its phonemic representation, and several types of information about the grapheme, such as the word it belongs to, the position of the grapheme in the word, the POS of the word, and the surrounding graphemes. These features are then given weights using techniques like gradient ascent. The features and their respective weights are then used to find the optimal sequence using dynamic programming algorithms like Viterbi [Sutton and McCallum, 2006].

Lastly, joint n-gram models use substring pairs of graphemes and phonemes so that the information about both part can be exploited [Jiampojamarn, 2011]. The graphemes and phonemes are first aligned so that each grapheme corresponds to its phoneme, using a Dynamic Time Wrapping (DTW) algorithm [Pagel et al., 1998]. Joint n-gram models can have orders ranging from 1 to 7. At run time, a simple search through the pairs would give the most probable sequence using the Viterbi algorithm again [Taylor, 2009].

3.5 Post-lexical processing

The results of G2P converters can be further processed to modify the generated pronunciations. The goal of post-lexical processing can be to improve the pronunciation or to model pronunciation for a given target expressivity. In this section, the main works in this area are reviewed.

[Miller, 1998] studied post-lexical phonology, which can lead to interspeaker variations, using neural networks. In order to predict the post-lexical pronunciation, the canonical pronunciations were encoded along with prosodic information and then fed into a neural network. Context of the phonemes were also fed into the neural network by using a window of three phonemes (one from left and one from right). For each target phoneme, the neural network outputted a post-lexical phoneme, a silence for deletions, or a diacritic to indicate minor changes in the pronunciation of the canonical phoneme. The system performed best when the variants of a phoneme were few, such as in case of word-initial vowel glottalization[2], where only two variants are available. However the systems struggled when the number of variants were more than two, such as the phoneme /t/.

In another study, [Vazirnezhad et al., 2009] followed a different technique which consisted of a decision tree and a contextual rule generator to produce post-lexical pronunciations. Given an input phoneme string, the decision tree was used to predict the phonemes that needed to be changed, while the contextual rules module was used to generate the post-lexical changes such as substitution, insertion or deletion of phonemes that were susceptible to change. The features that the authors included in the decision tree were rate of speech, word unigram probabilities, syllable location, and word stress. These features are known to have a huge impact on pronunciation variation in spontaneous speech. The authors compared their system to a baseline system which used only phonemic version of the words without using any of the aforementioned features. Their system achieved the best result when the lexicon contained 2.61 variants per word with 48% WER compared to the baseline system which had 54% WER. This shows that using extra features in addition to the sole phonemes is useful for post-lexical processing.

[Jande, 2003] studied how speech rate and speaking style affects pronunciation and phone-level reduction[3] in Swedish. The goal was to capture the general pronunciation variations in spontaneous speech rather than individual or dialect-related variations. Several rules were extracted to conduct the analysis such as haplology[4], forward and backward assimilation[5], and elision of the important phonemes such as /r/ and /h/. Common words matching these rules were then extracted, and

[2]Glottalization is the closure of vocal folds during the articulation of a sound.

[3]Phone-level reduction occurs when a word is uttered without pronouncing all the phonemes that it contains. Reduced words can still be understood. It exists mostly in spontaneous speech.

[4]Haplology is the deletion of successive identical syllables or consonant sound groups. For instance, the word *library* which has a canonical pronunciation of /laɪ.brər.i/ is pronounced as /laɪ.bri/ in spontaneous speech.

[5]Assimilation rules transfer a phonological feature or a set of features from a segment in a phonological string to the succeeding segment. For instance, the word *bags* in English is pronounced as /bægz/, and the /s/ is pronounced as /z/.

their realized post-lexical forms were compared to non-reduced (canonical) ones by presenting both to listeners in test utterances. The result of this evaluation showed that the canonical forms were considered more natural when the speech rate is low, while reduced forms were judged as the most natural when the speaking rate was medium or high.

Lastly, [Bennett and Black, 2003] tried to mimic an individual speaker by predicting a speaker's choice of pronunciation between reduced forms and full forms of English words. The focus was on the most common words which are known to have multiple pronunciations. For instance, the English words *the*, *a*, *to*, and *for* can have different pronunciations based on their context. One of the best known cases is the pronunciation of *the* which is mostly pronounced as /ðə/ when it is followed by a consonant-initial word and is pronounced as /ði/ when followed by a vowel-initial word. However, exceptions occur in many situations. For instance it might be pronounced as /ði/ even before consonant-initial words. For instance, when uttering the phrase *the car* in a context where the mentioned car is meant to be unique in some sense. The evaluation results on the words showed that the prediction for some of the words like *for* and *a* were mostly correct, while it failed to capture the variations in pronunciation of the other two words especially *to*.

Its very important to understand all the tasks and components which take place in pronunciation modelling and the way data is processed in each component. Therefore, this section presented a general overview of pronunciation modelling and the most important concepts in this area (G2P, post-lexical processing, features, etc.). In the next section another important area of phonology modelling will be discussed: disfluencies.

4 Disfluencies

One of the main differences between neutral speech and expressive speech is that the latter contains lots of disfluencies. Speech disfluencies are generally defined as phenomena that interrupt the flow of speech and do not add propositional contents to an utterance [Tree, 1995]. Disfluencies were considered as irregular phrases by most researchers and therefore had received very little attention from the researchers working in the area of phonology modelling. However, some studies have shown that there are actually remarkable regularities and patterns in disfluencies [Shriberg, 1994]. These regularities can be used to model disfluencies in order to improve expressive speech modelling.

Among researchers working in the area of disfluencies, there is an agreement on the structure of the surface form of disfluencies, that is what can be observed from them in a spoken utterance. [Shriberg, 1994] suggests some standard terms for the regions of disfluencies which is as in Figure 4. The section called Reparandum (RM) refers to the part of the linguistic message that will be deleted. Some researchers have preferred to refer RM to only the mistakenly uttered word such as "boston" in the above example as supposed to the entire deleted section. The Interruption Point (IP) is the exact place in which interruption occurs. That is when the speaker detects a trouble in his speech. The next section is Interregnum (IM) and refers to the section which is an indicator to the start of an editing phase or correction phase. Finally, Repair (RR) is the section in which the speaker corrects his/her speech.

Disfluencies can change form by language, but generally there are some accepted types which can be found in most languages. According to [Stolcke and Shriberg, 1996] the main disfluency types are:

- Filled pauses (FP): includes fillers like *uh* and *um*.
- Repetitions (REP): repeated words are considered as disfluencies, such as "*the* the word".
- Deletions (DEL): words without any correspondence such as "*I*... did you happen to see".

In this section, we will highlight the usage of some disfluency types and the possible positions in which they might appear.

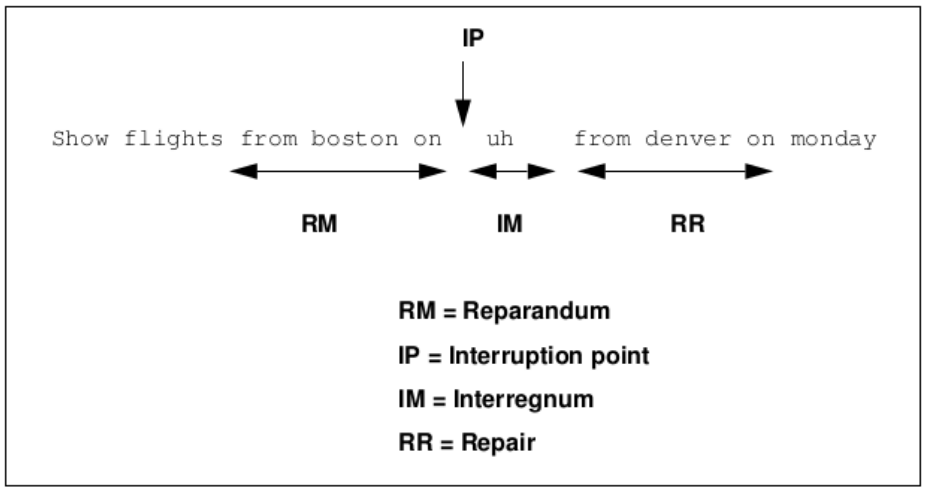


Figure 4: Terminology for disfluency regions (source: [Shriberg, 1994])

4.1 Usage and functions of disfluencies

Each of the above mentioned disfluency types have different functions and are used for different purposes. For example *uh* and *um* which by some researchers are considered to be *silent pauses* but filled without conveying any meaning [Eisler, 1968]. On the other hand, other researchers see these two disfluency types as *filled pauses* which are used to indicate the beginning of a delay that speakers use to search for a phrase and keep their conversation going on [Clark and Fox Tree, 2002].

Another pair of disfluencies that have similar characteristics but sometimes used for different purposes are *you know* and *I mean*. According to [Fox Tree and Schrock, 2002], these two markers have three basic meanings. The first is that these markers provide information about the speaker, including anxiety, uncertainty or lack of self confidence. They can also be an indicator of the speaker group or social class. The second proposal suggests that they are indicators for different situations such as informality or amusement. A great deal of these markers are found in informal speech and only few of them in formal speech. The final meaning is that they are used to express shared understanding on a topic, usually called positive politeness. This helps decrease the social distance between the speakers and makes speech more casual.

On the functions of the discourse marker *oh*, [Fox Tree and Schrock, 1999] points out that the function of *oh* might change based on the place it is found. However generally this discourse marker has functions like a sudden reaction to new or surprising information, a signal of an upcoming repair, and to demonstrate the speaker's engagement in the conversation (this is also sometimes referred to as backchannels). In addition *oh* signals an incoming separate piece of information that should be treated independently from the previous information, thus it helps listeners to understand more easily.

Regarding repetitions, they are considered as a sequence of processes that can be divided into four stages [Clark and Wasow, 1998]. In the first stage, the speaker makes the commitment to utter a phrase, and initiate it by saying the first word. In the second stage which occurs after the first word, the speaker suspends its speech because of different reasons such as thinking about reformulating his/her speech or cleaning throat. The third stage is between the suspension and the resumption point where the speaker utters a filled pause such as *uh* or remains silent. Lastly the speaker switched to what [Shriberg, 1994] calls *repair* and, in case of repetitions, it involves repeating the first word and then the correct phrase such as: "I uh I want to...". As for the function of repetitions, [Shriberg, 1994] reports they serve the smoothing process after a long pause.

4.2 Position of disfluencies

Disfluencies occur at different positions in sentence or more specifically at different positions of intonation units. The ratio of disfluencies might change based on the position and how much planning the speaker have done before the intonation unit. An analysis of the positions of *uh* and *um* has revealed three locations in the content of an utterance where these fillers usually appear [Clark and Fox Tree, 2002]. These locations are (i) at the boundary, (ii) after the first word, and (iii) at later positions of the intonation unit. Among these three locations, planning the speech is the most difficult at position (i), thus these fillers appear the most in this location. Logically, planning the speech is easier at (ii) and (iii), and less fillers appear there. The rate of occurrences of all disfluency types generally can be related to the level of planning of the speaker as pointed out by [Bortfeld et al., 2001], therefore these three positions can be generalized and applied to all other disfluencies.

In another experiment, [Swerts, 1998] analysed the distribution of filled pauses at different discourse boundaries[6]. The results showed that 78% of the phrases following a strong boundary had filled pauses and only 40% of the phrases following a weak boundary contained filled pauses. In addition, the filled pauses that follow a strong boundary mostly occur in initial position. Besides, the study showed that the existence of a boundary makes hesitations more likely. This finding is consistent with [Bell et al., 2003], suggesting that a large number of disfluencies occur at initial positions of an utterance.

Disfluencies have been considered as elements making the speech more difficult to comprehend for a long time. However some studies have reported that disfluencies actually help listeners to comprehend easier in many cases [Brennan and Schober, 2001, Fox Tree and Schrock, 1999]. This shows that studying disfluencies and understanding them can lead us to better model phonology and produce more natural and comprehensible synthetic speech utterances.

5 Conclusion

In this report, we reviewed expressivity and how it affects the variations in phonology. Expressivity is one of the most important factors that makes human speech natural. The three main declinations of expressivity are emotion, speaking style, and intention. They are known to have effects on phonology, i.e pronunciation and disfluencies. Expressivity can affect the acoustic side of speech, however in this study we only considered the linguistic effects of expressivity. Mainly, we discussed the area of pronunciation modelling which, in general, deals with finding the corresponding phonemic representations of a given text. Pronunciation modelling can also be adapted to model specific pronunciation patterns, variations or styles. The two respective key subtasks for this are G2P conversion and post-lexical processing. Pronunciation models can work with several types of input units like phonemes, syllables, and speech signals. Each of these units has its own pros and cons. Finally, disfluencies, which are an important characteristic of expressive speech, were discussed. Although disfluencies were considered as irregular events for a long time, some researchers have found regular patterns in disfluencies. These regularities can help us model disfluencies and to have better integration of disfluencies in phonology models.

The first step towards working on this subject will be extracting and analysing the characteristics of expressivity in speech corpora. As a preliminary work, we have analysed a spontaneous speech corpus named the Buckeye corpus and extracted some statistics about characteristics of spontaneous speech. The extracted information from the corpus confirms some of the findings of papers from this bibliographical review, such as the high ratio of phoneme deletion in the final syllables of

[6] A phrase is considered to be a discourse boundary if it can separate two pieces of information in the preceding and the following phrases. The degree of separation determines if a boundary is strong or weak. In cases where boundaries are determined by human evaluators, boundaries are said to be strong if the inter-annotator agreement is high, whereas the boundary is weak if there is a doubt about its presence.

words, and the reduction of phonemes in the frequent words. In the next step, we will try to build models that can learn the characteristics of expressive speech and rules which lead to variations in expressive speech, e.g, spontaneous speech. As an example, the model might learn rules such as the transformation of /t/ → /ʔ/[7] as in the word *that*. Next, we will try to integrate the model in a TTS system to generate expressive speech. Integrating such models in TTS might be more challenging than in ASR, because in ASR, the system might contain several alternative pronunciations of the same word without ranking them. Whereas in TTS the system has to generate one single output for each word, and in case of having multiple alternative pronunciations for the same word, they have to be ranked to determine the best one. Later, we will move to another important subject which is accents. To do so, first, we have to analyse an accented speech corpus, and then transfer our knowledge that we learnt before to accented speeches. Finally, in the last year of the PhD, disfluencies will be studied. The field of disfluency is a very challenging field, and it has been very poorly studied until now, especially in the domain of TTS.

References

[Adda-Decker et al., 2005] Adda-Decker, M., Boula de Mareüil, P., Adda, G., and Lamel, L. (2005). Investigating syllabic structures and their variation in spontaneous french. *Speech Communication*, 46(2):119–139.

[Adda-Decker and Lamel, 1999] Adda-Decker, M. and Lamel, L. (1999). Pronunciation variants across system configuration, language and speaking style. *Speech Communication*, 29(2–4):83–98.

[Arslan and Hansen, 1996] Arslan, L. M. and Hansen, J. H. (1996). Language accent classification in american english. *Speech Communication*, 18(4):353–367.

[Bell et al., 2009] Bell, A., Brenier, J. M., Gregory, M., Girand, C., and Jurafsky, D. (2009). Predictability effects on durations of content and function words in conversational english. *Journal of Memory and Language*, 60(1):92–111.

[Bell et al., 2003] Bell, A., Jurafsky, D., Fosler-Lussier, E., Girand, C., Gregory, M., and Gildea, D. (2003). Effects of disfluencies, predictability, and utterance position on word form variation in english conversation. *The Journal of the Acoustical Society of America*, 113(2):1001–1024.

[Bennett and Black, 2003] Bennett, C. L. and Black, A. W. (2003). Using acoustic models to choose pronunciation variations for synthetic voices. In *Proceedings of Interspeech*.

[Bortfeld et al., 2001] Bortfeld, H., Leon, S. D., Bloom, J. E., Schober, M. F., and Brennan, S. E. (2001). Disfluency rates in conversation: Effects of age, relationship, topic, role, and gender. *Language and speech*, 44(2):123–147.

[Brennan and Schober, 2001] Brennan, S. E. and Schober, M. F. (2001). How listeners compensate for disfluencies in spontaneous speech. *Journal of Memory and Language*, 44(2):274–296.

[Clark and Fox Tree, 2002] Clark, H. H. and Fox Tree, J. E. (2002). Using< i> uh</i> and< i> um</i> in spontaneous speaking. *Cognition*, 84(1):73–111.

[Clark and Wasow, 1998] Clark, H. H. and Wasow, T. (1998). Repeating words in spontaneous speech. *Cognitive psychology*, 37(3):201–242.

[Corley and Stewart, 2008] Corley, M. and Stewart, O. W. (2008). Hesitation disfluencies in spontaneous speech: The meaning of um. *Language and Linguistics Compass*, 2(4):589–602.

[Damper and Eastmond, 1997] Damper, R. I. and Eastmond, J. F. (1997). Pronunciation by analogy: Impact of implementational choices on performance. *Language and Speech*, 40(1):1–23.

[7] /ʔ/ is glottal stop.

[Dave, 2013] Dave, N. (2013). *Feature Extraction Methods LPC, PLP and MFCC In Speech Recognition.* ijaret.

[Dedina and Nusbaum, 1991] Dedina, M. J. and Nusbaum, H. C. (1991). PRONOUNCE: a program for pronunciation by analogy. *Computer Speech & Language*, 5(1):55–64.

[Deshpande, 2013] Deshpande, A. A. (2013). *Acoustic Data Based Grapheme to Phoneme Conversion.* PhD thesis, The Ohio State University.

[Dilts, 2013] Dilts, P. (2013). *Modelling phonetic reduction in a corpus of spoken English using Random Forests and Mixed- Effects Regression.* PhD thesis.

[Eide et al., 2004] Eide, E., Aaron, A., Bakis, R., Hamza, W., Picheny, M., and Pitrelli, J. (2004). A corpus-based approach to expressive speech synthesis. In *Fifth ISCA Workshop on Speech Synthesis.*

[Eisler, 1968] Eisler, F. G. (1968). *Psycholinguistics: Experiments in spontaneous speech.* Academic Press New York.

[Fant, 1971] Fant, G. (1971). *Acoustic theory of speech production: with calculations based on X-ray studies of Russian articulations*, volume 2. Walter de Gruyter.

[Fosler-Lussier, 1999a] Fosler-Lussier, E. (1999a). Contextual word and syllable pronunciation models. In *Proceedings of the 1999 IEEE ASRU Workshop.*

[Fosler-Lussier, 1999b] Fosler-Lussier, E. (1999b). Multi-level decision trees for static and dynamic pronunciation models. In *EUROSPEECH.*

[Fosler-Lussier and Morgan, 1999] Fosler-Lussier, E. and Morgan, N. (1999). Effects of speaking rate and word frequency on pronunciations in convertional speech. *Speech Communication*, 29(2):137–158.

[Fosler-Lussier, 1999c] Fosler-Lussier, J. E. (1999c). *Dynamic pronunciation models for automatic speech recognition.* PhD thesis, University of California, Berkeley Fall 1999.

[Fox Tree and Schrock, 1999] Fox Tree, J. E. and Schrock, J. C. (1999). Discourse markers in spontaneous speech: Oh what a difference an oh makes. *Journal of Memory and Language*, 40(2):280–295.

[Fox Tree and Schrock, 2002] Fox Tree, J. E. and Schrock, J. C. (2002). Basic meanings of< i> you know</i> and< i> i mean</i>. *Journal of Pragmatics*, 34(6):727–747.

[Govind and Prasanna, 2013] Govind, D. and Prasanna, S. M. (2013). Expressive speech synthesis: a review. *International Journal of Speech Technology*, 16(2):237–260.

[Govind and Prasanna, 2009] Govind, D. and Prasanna, S. R. M. (2009). Expressive speech synthesis using prosodic modification and dynamic time warping. *NCC 2009.*

[Greenberg, 1999] Greenberg, S. (1999). Speaking in shorthand–A syllable-centric perspective for understanding pronunciation variation. *Speech Communication*, 29(2):159–176.

[Guilleray, 2012] Guilleray, M. (2012). Towards a fluent electronic counterpart of the voice.

[Iida et al., 2003] Iida, A., Campbell, N., Higuchi, F., and Yasumura, M. (2003). A corpus-based speech synthesis system with emotion. *Speech Communication*, 40(1):161–187.

[Illina et al., 2011] Illina, I., Fohr, D., and Jouvet, D. (2011). Grapheme-to-Phoneme Conversion using Conditional Random Fields. In *Proceedings of Interspeech*, Florence, Italie. International Speech Communication Association (ISCA) et The Italian Regional SIG - AISV (Italian Speech Communication Association).

[International Phonetic Association, 1999] International Phonetic Association, editor (1999). *Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet*. Cambridge University Press.

[Jande, 2003] Jande, P.-A. (2003). Phonological reduction in swedish. In *Proceedings of ICPhS*, page 2557–2560.

[Jiampojamarn, 2011] Jiampojamarn, S. (2011). *Grapheme-to-phoneme conversion and its application to transliteration*. PhD thesis, University of Alberta.

[Karanasou, 2013] Karanasou, P. (2013). Phonemic variability and confusability in pronunciation modeling for automatic speech recognition.

[Laan, 1997] Laan, G. P. (1997). The contribution of intonation, segmental durations, and spectral features to the perception of a spontaneous and a read speaking style. *Speech Communication*, 22(1):43–65.

[Lafferty et al., 2001] Lafferty, J., McCallum, A., and Pereira, F. C. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

[Laver, 1994] Laver, J. (1994). *Principles of phonetics*. Cambridge University Press.

[Lopes and Perdigão, 2011] Lopes, C. and Perdigão, F. (2011). Phone recognition on the TIMIT database. *Speech Technologies/Book*, 1:285–302.

[Miller, 1998] Miller, C. (1998). Individuation of postlexical phonology for speech synthesis. In *The Third ESCA/COCOSDA Workshop (ETRW) on Speech Synthesis*.

[Moats, 2004] Moats, L. (2004). *LETRS, Language Essentials for Teachers of Reading and Spelling*. Sopris West Educational Services.

[Pagel et al., 1998] Pagel, V., Lenzo, K., and Black, A. (1998). Letter to sound rules for accented lexicon compression. *arXiv preprint cmp-lg/9808010*.

[Parlikar, 2013] Parlikar, A. (2013). *Style-Specific Phrasing in Speech Synthesis*. PhD thesis, Carnegie Mellon University.

[Pathak and Talukdar, 2013] Pathak, N. and Talukdar, P. H. (2013). The basic grapheme to phoneme (G2P) rules for bodo language. *International Journal*, 2(1).

[Polzin and Waibel, 1998] Polzin, T. S. and Waibel, A. (1998). Pronunciation variations in emotional speech. In *Modeling Pronunciation Variation for Automatic Speech Recognition*.

[Russell, 1980] Russell, J. A. (1980). A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161.

[Schachter and Shopen, 1985] Schachter, P. and Shopen, T. (1985). Parts-of-speech systems.

[Schröder, 2009] Schröder, M. (2009). Expressive speech synthesis: Past, present, and possible futures. In *Affective information processing*, page 111–126. Springer.

[Sejnowski and Rosenberg, 1987] Sejnowski, T. J. and Rosenberg, C. R. (1987). Parallel networks that learn to pronounce english text. *Complex systems*, 1(1):145–168.

[Shriberg, 1994] Shriberg, E. E. (1994). *Preliminaries to a theory of speech disfluencies*. PhD thesis, University of California.

[Stolcke and Shriberg, 1996] Stolcke, A. and Shriberg, E. (1996). Statistical language modeling for speech disfluencies. In *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*, volume 1, page 405–408. IEEE.

[Strik and Cucchiarini, 1999] Strik, H. and Cucchiarini, C. (1999). Modeling pronunciation variation for ASR: a survey of the literature. *Speech Communication*, 29(2–4):225–246.

[Sutton and McCallum, 2006] Sutton, C. and McCallum, A. (2006). An introduction to conditional random fields for relational learning. *Introduction to statistical relational learning*, page 93–128.

[Swerts, 1998] Swerts, M. (1998). Filled pauses as markers of discourse structure. *Journal of pragmatics*, 30(4):485–496.

[Taylor, 2005] Taylor, P. (2005). Hidden markov models for grapheme to phoneme conversion. In *Proceedings of Interspeech*, page 1973–1976.

[Taylor, 2009] Taylor, P. (2009). *Text-to-speech synthesis*. Cambridge University Press.

[Tellier et al., 2010] Tellier, I., Eshkol, I., Taalab, S., Prost, J.-P., et al. (2010). Pos-tagging for oral texts with crf and category decomposition. *Research in Computing Science*, 46:79–90.

[Tree, 1995] Tree, J. E. F. (1995). The effects of false starts and repetitions on the processing of subsequent words in spontaneous speech. *Journal of memory and language*, 34(6):709–738.

[Van Den Bosch and Daelemans, 1993] Van Den Bosch, A. and Daelemans, W. (1993). Data-oriented methods for grapheme-to-phoneme conversion. In *Proceedings of the sixth conference on European chapter of the Association for Computational Linguistics*, page 45–53. Association for Computational Linguistics.

[Vazirnezhad et al., 2009] Vazirnezhad, B., Almasganj, F., and Ahadi, S. M. (2009). Hybrid statistical pronunciation models designed to be trained by a medium-size corpus. *Computer Speech & Language*, 23(1):1–24.

[Viterbi, 1967] Viterbi, A. J. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Trans. on Information Theory*, 13(2):260– 269.

[Wang and King, 2011] Wang, D. and King, S. (2011). Letter-to-sound pronunciation prediction using conditional random fields. *Signal Processing Letters, IEEE*, 18(2):122–125.

[Wieling et al., 2009] Wieling, M., Prokić, J., and Nerbonne, J. (2009). Evaluating the pairwise string alignment of pronunciations. In *Proceedings of the EACL 2009 Workshop on Language Technology and Resources for Cultural Heritage, Social Sciences, Humanities, and Education*, page 26–34.